

Frontiers in Clinical Science

Challenges of Scientific Data Management for Large Epidemiologic Studies

M.K. HENDERSON,¹ C. MOHLA,¹ K.B. JACOBS,² and J.B. VAUGHT¹

ABSTRACT

The U.S. National Cancer Institute's Division of Cancer Epidemiology and Genetics (DCEG) conducts population-based and interdisciplinary research to discover the genetic and environmental determinants of cancer. Many DCEG studies are large, multi-institutional, and long-term with national and international study sites involved in the multiple research steps. Current information technology challenges involved in such epidemiological studies include: (1) management and harmonization of a multitude of data types (demographic, environmental, biospecimen, laboratory, analytic, molecular, etc.); (2) unprecedented amounts of data; (3) efficient data mining to derive insights into disease etiology; and (4) secure collaboration between study management systems. If not adequately addressed, all of these challenges will increase the cost of performing studies and decrease the speed of publication. DCEG is examining current data management practices to better utilize recent advances in information technology to enhance its scientific program. This analysis is providing strategic guidance in enhancing interoperability among current data systems, further automating specimen management practices, defining metadata strategies to allow for better cross study comparability and reusability, and in planning for integration of new technologies in support of DCEG's epidemiology research. Early results from the effort include better communication of information technology requirements between contractors and investigators, as well as progress on several focused data interoperability projects, including Web services transactions for biorepository interoperability and improved analytic support utilizing data warehouses.

INTRODUCTION

THE U.S. NATIONAL CANCER INSTITUTE'S (NCI) Division of Cancer Epidemiology and Genetics (DCEG) conducts population-based and interdisciplinary research to discover the genetic and environmental determinants of cancer. Over the past 20 years, epidemiology at the NCI has grown from a cottage industry to a full division with nearly 200 scientists engaged in 350 active research studies. Epidemiologic studies are multicenter, multidisciplinary case-control and cohort studies. The majority of

DCEG's epidemiologic studies utilize molecular technology to examine gene-environment interactions, involving the collection and processing of biospecimens for high-throughput genotyping, proteomic analysis, and other laboratory applications. The successful management of such large population-based studies requires the efficient handling of biospecimens and demographic data from collection to storage to analysis.

The DCEG biospecimen program currently stores almost 11 million biological specimens, including over 122 different specimen types,

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Rockville, Maryland.

²Capital Information Technology Services, Inc., Rockville, Maryland.

and is among the largest specimen collections at the National Institutes of Health (NIH). The DCEG specimen collections have grown over nine-fold since 1988, and, since 1999, at an annual rate of over 15%. The laboratories and biorepositories that support DCEG's molecular epidemiology program collect and process biological specimens for studies that range from less than 100 study participants in small case-control or family studies to over 10,000 participants in some prospective cohort studies. Almost all of the studies involve the collection of multiple biologic samples, contributing to the explosive growth in DCEG's specimen collections and to the accumulation of data that result from processing and analyzing the specimens.

DCEG investigators continue to explore new laboratory techniques that include serum proteomics profiles, protein expression arrays, and tissue microarrays. Implementation of new techniques and technologies in large molecular epidemiology studies generate massive amounts of data necessitating a scalable, high-performance data management system that will allow:

- Access to heterogeneous data and tools.
- Integration of a variety of data types.
- Management of vast quantities of data.
- Data mining to detect underlying patterns.
- Secure collaboration.
- Standardized and *ad hoc* query.

Currently available data management and analysis tools are not capable of meeting the challenges posed by such large-scale requirements, in terms of performance, scalability, and user-friendly interfaces. Also, large-scale computing requires new approaches to solving the problems of storing, retrieving, managing, sharing, visualizing, organizing, and analyzing data. As a global research enterprise, cancer epidemiology is in need of support for secure sharing and collaboration for investigator teams (1).

To evaluate its data management needs, DCEG conducted an internal bioinformatics assessment. DCEG researchers are developing new databases for each study, because various types of scientific data are stored in disparate formats, in disparate systems, at multiple sites,

and with different workflows. These characteristics make it difficult to merge and analyze data, as well as to aggregate data across studies for meta-analyses. Most of the data systems do not require or allow real-time access to data collection. Fine-tuning the data collection strategy for a study in the field collection stage is virtually impossible, making the retrieval of missing data problematic. To support the participation in several cancer consortia, DCEG researchers need to perform cross-study comparisons to increase study sizes and provide validation of published findings.

INFORMATICS STRATEGY

As a result of the internal bioinformatics assessment, DCEG has developed a three-tiered informatics strategy. In the short term, the focus will be to increase communications between research teams. Plone, an open-source content management and collaboration system, is being implemented to manage study communications and document preparation interactions (2). In addition, communications among the research field stations, laboratories, and biorepositories are being enhanced. In the medium term, Web services are being used to automate communications between the biospecimen inventory system and the laboratory information management systems (Fig. 1). This will reduce cost and enhance legacy systems by making them more interoperable. The long-term plan is to build the data management infrastructure as a Core Data Architecture (CDA). The CDA would consolidate data management systems and application across many DCEG research initiatives and be used by DCEG researchers, external collaborators, and contractors. The interfaces for each study could be customized for specific study applications, but the infrastructure would stay constant.

A DATA WAREHOUSE APPROACH

The proposed CDA is a centralized scientific data management system and a knowledge-based architecture for mining large datasets. The primary objective of the CDA is to bring

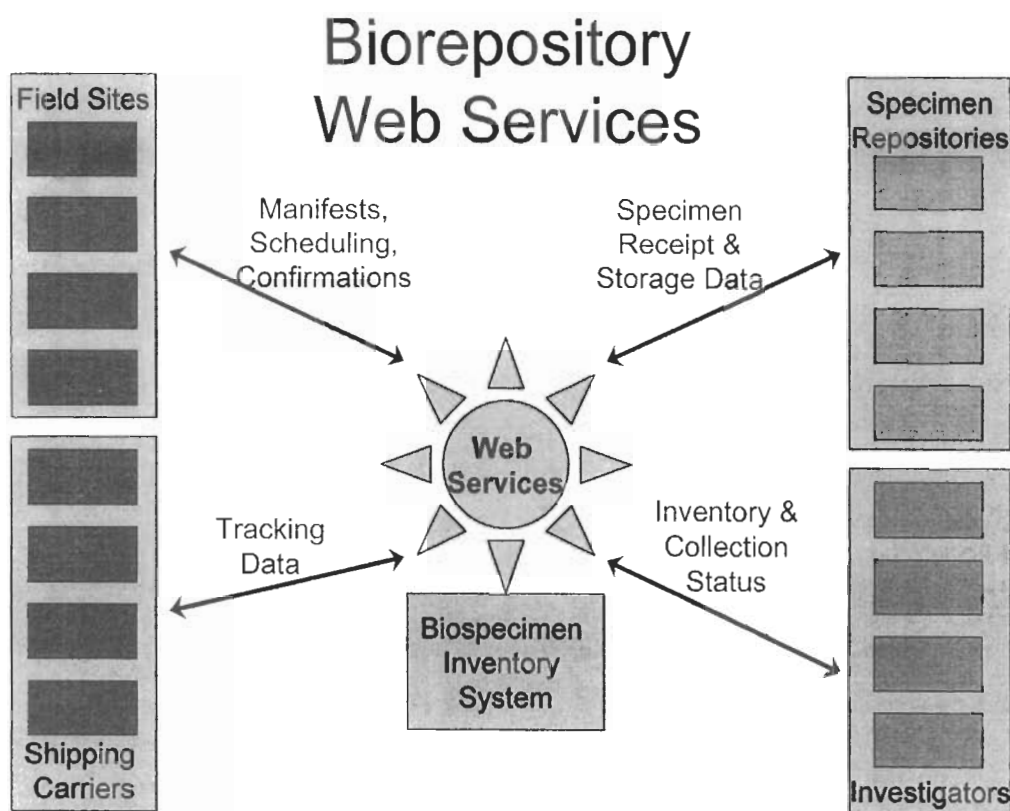


FIG. 1. Automated communications through Web services.

together information from disparate sources and put the information into a uniform format that is conducive to analysis and archival storage.

CENTRAL DATA WAREHOUSE STRUCTURAL OVERVIEW

Structurally, the data within the CDA would flow from many study and laboratory databases (feeder systems) (Fig. 2). Although many of these feeder systems utilize disparate data models, encoding standards, and transfer mechanisms, the "Data Load and Staging Area" will take those inputs from these various and incompatible systems. The appropriate transformations and data mappings will be performed to harmonize the data to conform with standardized data element definitions. The transformed data and metadata will then be loaded into a Central Data Warehouse (CDW), from which it can be utilized by a variety of applications.

CDA COMPONENTS

The preliminary list of architectural components needed to build the CDA includes:

- CDW in which all DCEG study data and metadata will be stored. Study managers will employ the infrastructure of the data warehouse to store master files, raw data, and derived variables. Other study metadata will also be accommodated, such as decision logs, raw data sets, version control information, and edit check reports.
- Metadata Management Interface will describe and link data stored in the warehouse to the necessary metadata definitions needed to interpret the data; e.g., coding sheets, data dictionaries, derived variable definitions, and questionnaire modules.
- Data Collection Management System will be used to track data entry, apply edit checks, log discrepancy resolutions, maintain audit logs, and track the progress of data collection and cleaning efforts.

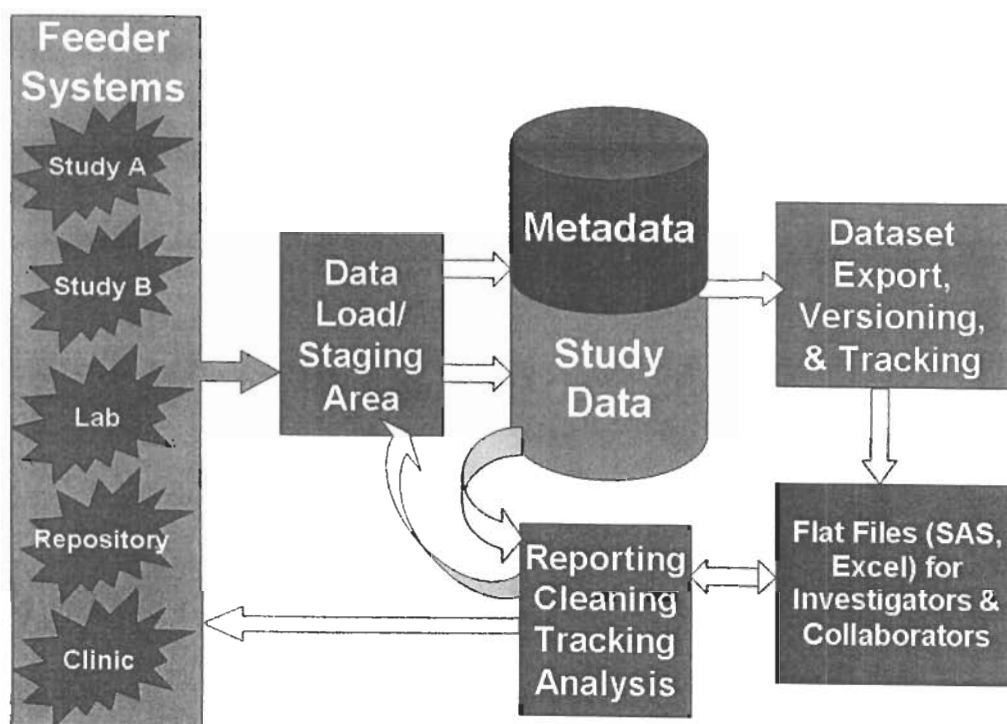


FIG. 2. CDA data flow.

- Data Mapping and Transformation System will translate multiple data standards. Applications include translating data between coding standards (e.g. translate between ICD-9, ICD-10, and ICD-0), between metadata versions, and export DCEG data for use by outside groups.
- Dataset Export and Analysis management system will track data releases, data versions, and flexible data reporting. It will utilize the data warehouse, metadata system, and data transformation system to provide version-controlled master files and subset files for analysis. It will also facilitate *ad hoc* queries, bulk data extraction, and data reporting.
- Security and Administration Module will manage users, roles, permissions, data access policies, system configuration, and other administrative tasks.

Additional interface tools will be needed, including those for specialized data mining, advanced reporting tools, data collection, and analytic applications. However, all of these can use the underlying architecture to access the data and feed results back into the warehouse.

INTEGRATED METADATA

One of the most important characteristics of the proposed CDA is the tight integration of data with the relevant metadata needed to interpret it. This critical feature distinguishes the CDA approach from most existing commercial and open-source software packages currently available. The goal of developing metadata is to enable semantic interoperability—the ability to represent information precisely enough that it may pass between humans and electronic representations precisely without requiring absolute central control of data systems or external human expertise.

The NCI Center for Bioinformatics has begun the development of the cancer Biomedical Informatics Grid or caBIG (3). caBIG is a voluntary network or grid connecting individuals and institutions to enable the sharing of data and tools, with a goal of creating a World Wide Web of cancer research. The focus is to speed the delivery of innovative approaches for the prevention and treatment of cancer. The infrastructure and tools created by caBIG also have broad utility outside the cancer community. It is the caBIG

cancer data standards repository (caDSR) that DCEG, and the extramural population sciences community, will use to build and maintain a repository of common data elements for standardization of terms and data storage practices.

IMPLEMENTATION

The CDA will be built in phases and will ultimately comprise a series of general and reusable modules, and data management applications that will include data cleaning, visualization, reporting, exports, version control, release tracking, study design tools, *etc.* Whenever feasible, the development will employ free/open software components and open standards for data representation, data storage, and data transmission. The implementation will be compliant with the caBIG development initiative standards and will leverage the caBIG open-source APIs and tools for population sciences research.

SUMMARY

The goals of DCEG's data management improvements are to speed discovery of research, lower study costs, and develop reusable, open

source and open standards-based data systems for new research endeavors.

ACKNOWLEDGMENTS

The authors would like to acknowledge the members of the Information Technology Oversight Committee for their dedicated effort in the analysis and recommendations for this project.

REFERENCES

1. Choudhary A, Taylor V, et.al. High-Performance Data Management, Access, and Storage for Tera-Scale Scientific Applications: project, sponsored by the Department of Energy's (DOE) Accelerated Strategic Computing Initiative (ASCI), 1998–2001.
2. Plone (<http://www.plone.org/>).
3. NCI's cancer Biomedical Informatics Grid, caBIG (<http://cabig.nci.nih.gov/>).

Address reprint requests to:
Marianne K. Henderson, M.S.

Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH, DHHS
6120 Executive Boulevard, MSC 7242
Rockville, MD 20892–7242

E-mail: mk149c@nih.gov

